Published online 31 May 2013 *Nucleic Acids Research, 2013, Vol. 41, Web Server issue* ***W441–W447***
doi:10.1093/nar/gkt428

PlantLoc: an accurate web server for predicting plant protein subcellular localization by substantiality motif

Shengnan Tang[1], Tonghua Li[1,*], Peisheng Cong[1], Wenwei Xiong[2], Zhiheng Wang[1] and Jiangming Sun[3]

[1]Department of Chemistry, Tongji University, Shanghai 200092, China, [2]Department of Biology and Molecular Biology, Montclair State University, Montclair, NJ 07043, USA and [3]Units of Molecular Metabolism, Lund University Diabetes Centre, SE-205 02 Malmö, Sweden

Received February 8, 2013; Revised April 18, 2013; Accepted April 27, 2013

ABSTRACT

Knowledge of subcellular localizations (SCLs) of plant proteins relates to their functions and aids in understanding the regulation of biological processes at the cellular level. We present PlantLoc, a highly accurate and fast webserver for predicting the multi-label SCLs of plant proteins. The PlantLoc server has two innovative characters: building localization motif libraries by a recursive method without alignment and Gene Ontology information; and establishing simple architecture for rapidly and accurately identifying plant protein SCLs without a machine learning algorithm. PlantLoc provides predicted SCLs results, confidence estimates and which is the substantiality motif and where it is located on the sequence. PlantLoc achieved the highest accuracy (overall accuracy of 80.8%) of identification of plant protein SCLs as benchmarked by using a new test dataset compared other plant SCL prediction webservers. The ability of PlantLoc to predict multiple sites was also significantly higher than for any other webserver. The predicted substantiality motifs of queries also have great potential for analysis of relationships with protein functional regions. The PlantLoc server is available at http://cal.tongji.edu.cn/PlantLoc/.

INTRODUCTION

Because the subcellular localization (SCL) of protein is highly correlated with its function, interaction partners and biological processes, it is a hot research topic in biology. According to the UniportKB database released on March 2012 (1,2), the original viridiplantae protein entries were 1 027 477. However, only 15 792 entries have any experimentally annotated SCL. It is both time consuming and expensive to determine the localization of a new protein using experimental methods. Computational prediction of SCLs has become a necessary alternative (3,4).

In recent years, various prediction methods have been developed to predict protein SCLs. These approaches may be classified into different categories based on exploiting difference features as follows. (i) The feature is generated on sequence information, such as amino acid composition (5–10), N-terminal sequence (11), pseudo-amino acid composition (5,12) and PSSM (position-specific scoring matrix). (ii) The feature is generated by making use of Gene Ontology (GO) annotations (13–15), or textual information (15) from Swiss-Port keywords to predict SCL. (iii) The feature is generated by hybrid methods, which usually combine sequence information and annotation information (16–20). Despite these features playing important roles in prediction the accuracy of prediction still needs improvement, especially for plant proteins (21–23). Additionally, the complexities of predictive models are difficult for users to understand why a prediction was made.

In the past, the concept of the localization motif (LM) had been proposed (24). Recently a novel feature, the localization motif, was proposed by our group (25,26). An LM was defined as a gapped or ungapped fragment of amino acids that were a conserved pattern in a subcellular domain and existed in the N-terminus peptides of sequences. We confirmed that LMs could be utilized as features to accurately predict protein SCLs by using support vector machine, and it is possible to directly use it to predict SCL without other information.

Here, we present PlantLoc, a highly accurate and fast web server for predicting the multiple-site SCLs of plant

*To whom correspondence should be addressed. Tel: +86 21 65983987 Fax: +86 21 65983987; Email: lith@tongji.edu.cn
Present address:
Tonghua Li, Chemistry Department, Tongji University, Siping Road No.1239, Shanghai 200092, China.

proteins from sequences without any annotation information or any machine learning algorithm. It provided predicted SCL results, confidence estimates and which substantiality motifs and where they located on the query sequences. PlantLoc generated the LMs by using training dataset (4412 entries). The obtained LMs constituted 11 libraries for 11 SCLs of plant proteins. According to the hit numbers of all LM libraries, PlantLoc gave the probabilities of the query sequence for each localization domains. Compared with six plant SCL prediction webservers, by using a new benchmark test dataset of 230 entries, the overall accuracy was 80.8% [cf. iLoc-Plant (21): 22.2%; mGOASVM (19): 51.3%; WegoLoc (17): 57.0%; YLoc (20): 47.8%; ngLOC (18): 26.5%; and WoLF PSORT (27): 61.7%]. Additionally, we tested PlantLoc's ability to predict multiple localizations. PlantLoc reliably predicted the proteins with multiple localizations and outperformed the best predictors in this area. Moreover, the obtained substantiality motifs are the conservative patterns in protein sequence which can facilitate users for understand why a prediction was made and further biological function analysis.

MATERIALS AND METHODS

Protein sequences were collected from the UniProtKB/Swiss-Prot protein knowledgebase (http://www.uniprot.org/) according to the annotation information in the CC (comment or notes) and OC (organism classification) fields. Some proteins may simultaneously exist in two or more SCLs. Training datasets were collected from the UniProtKB/Swiss-Prot release of March 2012. Plant proteins can be localized in the chloroplast (CHL), cell wall (CEL), cytoplasm (CYT), endoplasmic reticulum (END), extracellular space (EXC), Golgi apparatus (GOL), mitochondrion (MIT), nucleus (NUC), peroxisome (PER), plasma membrane (PLA) and vacuole (VAC). To reduce homology bias, a redundancy cutoff was operated by a culling program CD-HIT (http://weizhong-lab.ucsd.edu/cdhit_suite/cgi-bin/index.cgi) to window those sequences, which have 60% sequence identity) to any other in the same SCL (Figure 1). A test set (S^{Test}) was collected from UniProtKB released from March 2012 to October 2012 and reduced the redundancy to 60% refer to the training dataset. The total number of different proteins in S^{Test} was 206 and the total number of locations was 230. A multiple-site dataset called S^{MS} included 24 proteins with 51 localizations.

In Figure 1, the experimentally training (14 401 entries), reduced redundancy training (4436, the sum of each number in cylinder) and testing (230 entries) sets are divided by time (bottom). The names and the numbers of sequences of 11 subcellular domains of plant proteins are shown in different colors (top). The 36 651 entries (bottom middle) obtained by 'similarity' annotation were used in selection of LM (see text).

A substantiality motif was defined as assembled fragments of sequences which were interpretable characters of SCL. A substantiality motif for a query sequence was generated by assembling some LMs. These LMs hit the query and just liked bricks to construct flexible substantiality motifs. There were three steps to obtain a substantiality motif. In the first step, the LM program was carried out to generate candidates of LMs from the training dataset S^{Train}. N-terminal sequence information has been widely used for predicting SCL (7,28–30). In previous studies we also found that most of the motifs were positioned near the N-terminus of protein sequences, where signal peptides were generally considered to be present (26). Please see detail information in Supplementary Material (A). If the length of a sequence is less than 200 amino acids, all sequence will be used for extraction of LM candidates. If the length of a sequence is more than 200 amino acids, only 200 residues of N-terminus will be considered for the extraction of candidates. In previous studies, an algorithm that could extract local combinational variables with fixed locations from aligned sequence was successfully used to predict DNA binding and protein shape string (31,32). Here, the algorithm was developed to extract LM candidates from unequal-length sequences at different locations. There were two parameters in this procedure, the found number threshold and the span. The found number is the number of times a given LM candidate was present in a subcellular domain. If the found number threshold is low the coverage of training sequences will be high and the numbers of LM candidates will be very large. The span is the number of gaps between two residues that is used in generation of 2-length seeds (Figure 2). The larger the span, the more 2-length seeds are generated. All 2-length seeds with frequencies greater than or equal to the given found number threshold were enumerated. Then 2-length seeds were merged according to the same prefix-of-seed and suffix-of-seed (marked in Figure 2, left), and 3-letter seeds were generated. The new 3-length seed will survive if its frequency is greater than or equal to the found number threshold. This iterative circulation was carried out until no seeds survived. Finally, all surviving seeds were collected as LM candidates.

In the second step, LMs were selected from millions of generated candidates. For a subcellular domain, an LM is determined as belonging to this domain if the LM only matches sequences in this domain and does not match any sequences in all other domains. In order to reduce false positive rates, which are the inherent weaknesses of customary motif discovery algorithms, we enlarged the dataset and especially the negative set. The datasets used in this step included not only those of proteins annotated 'experimentally' SCL (14 401), but also from those annotated with 'by similarity' (36 651 entries of 11 SCLs from UniProtKB released on March 2012. For each subcellular domain the sequences belonging to this domain were considered as the positive set and the sequences of all other domains were considered as the negative set (Figure 2, middle). Thus the number of LMs was lowered and the false positive rates were greatly reduced. After selection, the LMs and their frequencies from the training sets for a special SCL were constituted in an LM library (Figure 2, right). There were 11 libraries for plant proteins.

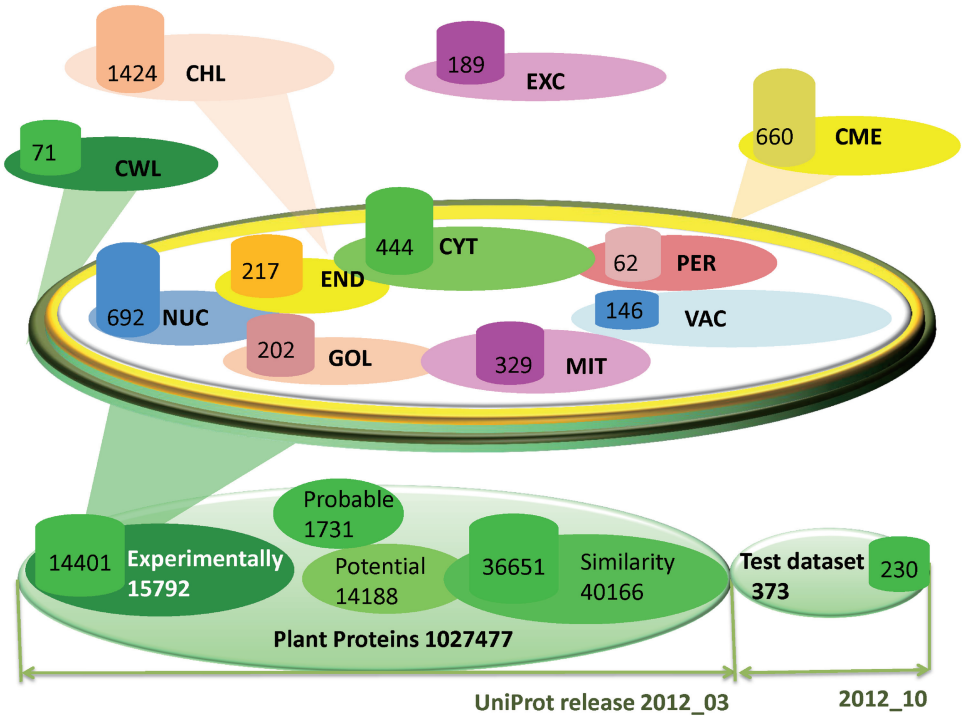

Figure 1. The numbers of plant proteins for the training and testing datasets. The S^{Train} (4436 entries, the sum of each number in cylinder) and S^{Test} (230 entries) sets are divided by time (bottom). The names and the numbers of sequences of 11 subcellular domains of plant proteins are shown in different colors (top). The 36 651 entries (bottom middle) obtained by 'similarity' annotation were used in selection of LM (see text).

In the third step, a query sequence was identified as belonging to a subcellular domain according to the hit numbers of LMs in LM libraries. When a library had the highest hit number, the query was identified as belonging to this domain. When more than one library had equal highest hit numbers and the hit numbers were above a threshold (say 10) the query was identified as of multiple sites. The substantiality motif of a query was assembled of hit LMs of a special domain and was shown in the output (Figure 2, bottom). When there were >10 LMs hit, only the 10 LMs with the highest frequencies were shown. In the output section of PlantLoc, the hit number is expressed as relative probability. If the hit number achieves the threshold, the probability is defined as 100%. Our approach shows which substantiality motifs and where they locate on the query sequences and this has great potential for analysing relationships with protein functional regions.

RESULTS

For PlantLoc the threshold number was set in the range of 0–10 according to the numbers of proteins in the training subcellular domains. When a subcellular domain had <100 proteins, the threshold was set as 2. The number of span was set to 0–10. We tested the performance of PlantLoc on training datasets. We tested PlantLoc's ability to predict multiple localization sites with 5-fold cross-validation [Supplementary Material (B)]. The overall accuracy of the S^{Train} was 96.3%. Because the LM was generated and selected from the training dataset, it was easy to understand why it achieved such high accuracy.

We tested the performance of PlantLoc on the independent dataset S^{Test} and compared PlantLoc with six other SCL predictors (iLoc-Plant (21), mGOASVM (19), WegoLoc (17), Y-LOC (20), ngLOC (18) and WoLF PSORT (27)) based on homology-based method, motif method, GO method. WoLFPSORT (27) converted protein amino acid sequences into numerical localization features, such as sorting signals, amino acid composition and functional motifs. YLoc (20) derived a lot of features from amino acid composition and pseudo composition. In addition, it included PROSITE motifs and GO terms from close homologies. iLoc-Plant (21) proposed by Chou group used PSSM, GO and sequential evolution. WegoLoc (17) was a homology-based and weighted GO-based approach. mGOASVM (19) also used homologous and GO information. ngLOC (18) was developed on fixed-length peptide, called *n*-gram. The individual prediction performance was evaluated using the overall accuracy. Except for the WoLF PSORT webserver, the others provided only the one or two predicted SCLs. Some webservers such as ngLoc and WoLF PSORT provided two or more probable SCLs. The evaluation results are summarized in Figure 3. For WoLF PSORT, the first and second predicted SCLs are defined as predicted SCLs. The overall accuracy of PlantLoc was 80.8%, which is much higher than by the other methods.

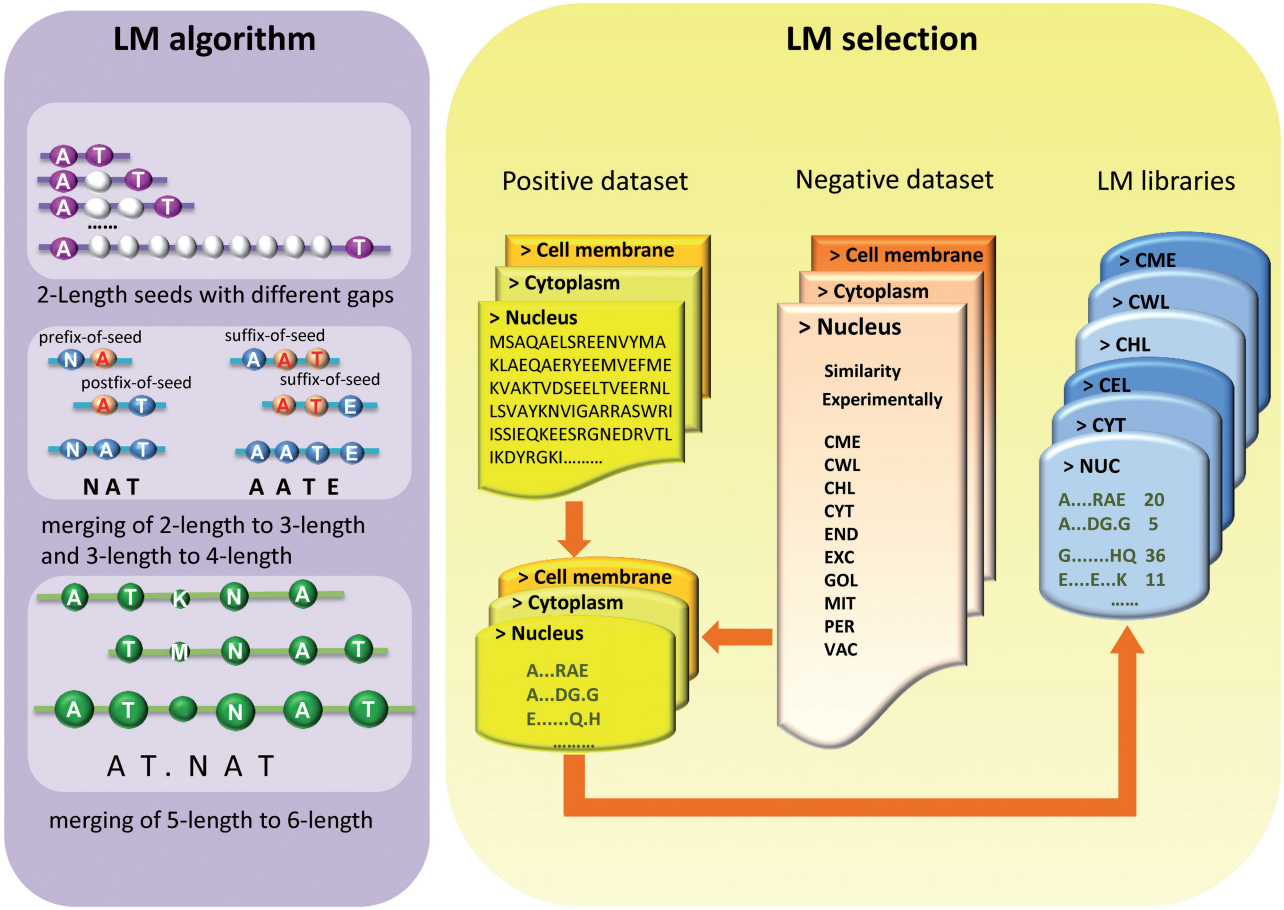

Figure 2. The process of the LM library building strategy. The LM algorithm is an enumeration and merging procedure, prefix-of-seed and suffix-of-seed are marked with tangerine. In LM selection, the negative set is enlarged. The LM libraries contain LMs (expressed by characters) and their frequencies in training sets.

Moreover, the results of predicted and probable SCLs were also evaluated as follows: WoLF PSORT 74.3%, ngLOC 59.6% and PlantLoc 90.4%. In Figure 3, the results show that PlantLoc performed markedly better than the existing predictors that based on motif, homology and GO information method with the capacity to deal with a multi-label plant protein.

We tested PlantLoc's ability to predict multiple localization sites on S^{MS} and compared the performances with six other webservers (Table 1). PlantLoc performed markedly better than any of the existing predictors with the capacity to deal with a multi-label plant protein. We defined the multiple-site accuracy (MSA) and the ratio of accuracy (RA) to assess the accuracy for predicting multiple sites, as follows:

$$\mathrm{MSA} = \mathrm{NCPS}/\mathrm{TNRMS} \quad (1)$$

$$\mathrm{RA} = \mathrm{NCPS}/\mathrm{TNPS}, \quad (2)$$

where NCPS is the number of correct predicted sites, TNRMS is the total number of real multiple sites and TNPS is the total number of predicted sites.

WegoLoc provided the confidence estimates of each location. For statistical reasons, the first three localizations were defined as its predicting result. For the other five webservers, all prediction results were defined as predicting result. For PlantLoc, MSA = 86.3% and RA = 100.0%. The ability of PlantLoc to predict multiple sites was better than other six webservers.

WEBSERVER

The PlantLoc server is free available at http://cal.tongji.edu.cn/PlantLoc/. The PlantLoc webserver runs on a Windows 64 server of 2.0 GHz Intel Xeon processors that consists of four cores. It is composed of a front-end web application and a back-end execution cluster. The front-end is written in Java and Java Server Page and uses the Microsoft SQL Server database. The LM software was developed in C#, and can be freely downloaded. With the help of the template Perl program, it is easier to submit sequences and parse the result for users.

Input description

Users can easily input a FASTA format protein sequence in the textbox or FASTA format file (up to 50 entries). Users can then bookmark 'MY TASK' page (about 30 s

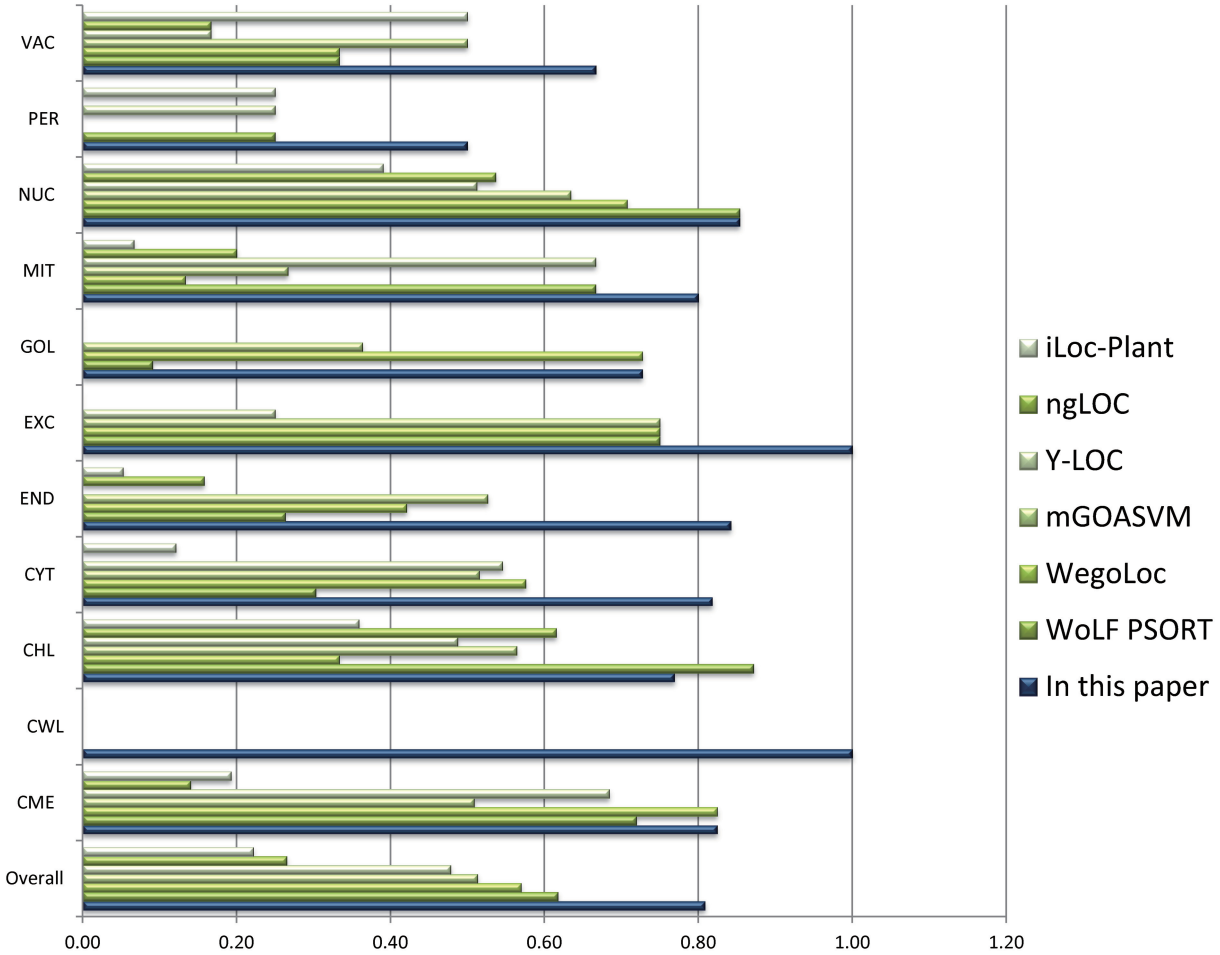

Figure 3. Compared results with other methods on S^{Test}.

Table 1. Performance of PlantLoc and other webserver on S^{MS}

	PlanLoc	iLoc-Plant	Yloc	mGoasum	ngLOC	WoLF PSORT	WegoLoc
MSA (%)	86.3	25.5	35.3	41.2	39.2	58.8	64.7
RA (%)	100.0	50.0	75.0	72.4	27.8	42.9	45.8

per sequence) and access at a later time. If a user provides an Email address (optional), the address will be considered as an ID to retrieve the results of all tasks the user has ever submitted. For large-scale predictions, the standalone program for finding LMs can also be downloaded from the webserver.

Output description

All prediction results provided a graphic representation of the probabilities of predicted SCLs (Figure 4). The identified SCLs, probable SCLs and the substantiality motifs, including their position on the sequence, are shown in Figure 4. When substantiality motifs were determined in the query sequences, they were represented as amino acid characters. The rest of positions were represented as dots. All the substantiality motifs can be download by pressing 'download' in the result file.

The substantiality motifs are very important since they tell the user why the proteins are predicted in this localization. In Figure 5, there is an example of potential relationship between substantiality motifs and functional regions annotated by UniportKB. The substantiality motifs (V.ALN. . . . L, KYCG. . Y. GCP. E. PCD. . D. CC) were localized on the signal peptide and the metal-binding annotation region of the sp_Q8S8N6_PIA2A_ARATH_Phospho sequence. The substantiality motifs may have relationship with signal peptide and functional regions which is the reason why it can provide more accuracy result. Moreover, if the users found the substantiality motif, they can also use UniportKB database to find potential functional regions for deep research.

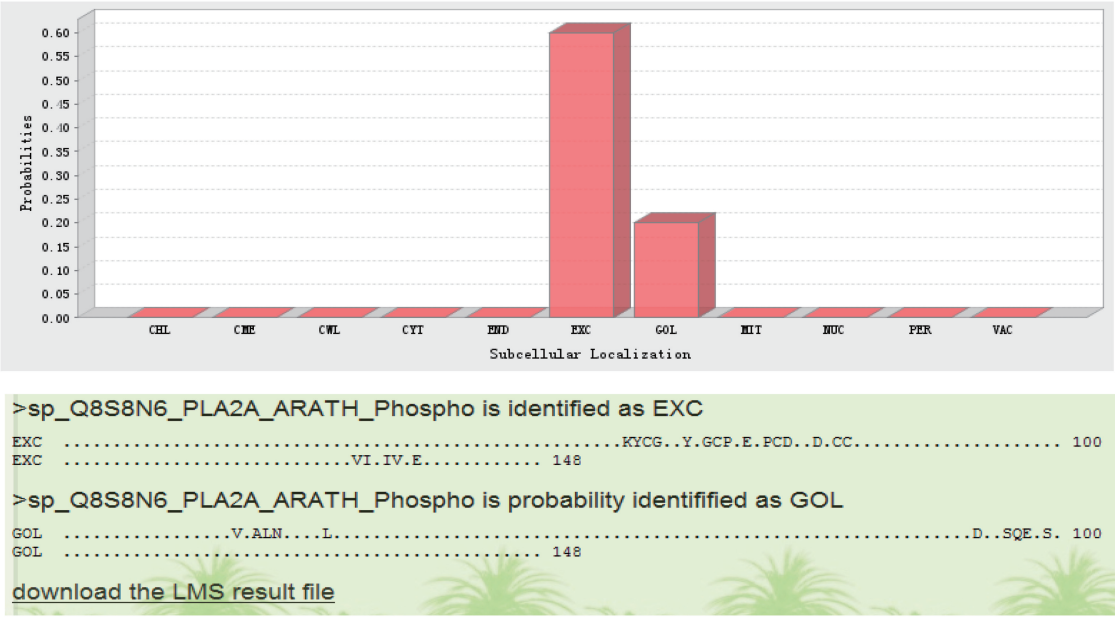

Figure 4. A screenshot of a PlantLoc output and obtained substantiality motifs of Q8S8N6 (protein ID). The probability of prediction expressed by graph. The identified SCL(s) and probability localization(s).

Figure 5. The substantiality motifs with annotation by UniProtKB. Characters colored blue are annotated from UniProtKB. Characters colored red are substantiality motifs for EXC and GOL.

DISCUSSION

The PlantLoc server has two innovative characters: building LM libraries by recursive method without alignment and GO information and establishing simple architecture for rapidly and accurately identifying plant protein SCLs without a machine learning algorithm. In contrast to other webservers, PlantLoc performs excellently not only on single localizations but also on multiple-site proteins. The substantiality motifs can explain why a prediction is made and which substantiality motifs are responsible for prediction. The substantiality motifs will be very important for users on further functional analysis and can be applied to a wide range of sequence identities and so provide a practical tool for biologists.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors are grateful to the anonymous reviewers for their valuable comments and help in improving the manuscript.

FUNDING

National Natural Science Foundation of China (NSFC) [20705024, 21275108]. Funding for open access charge: NSFC.

Conflict of interest statement. None declared.

REFERENCES

1. The UniProt Consortium. (2013) Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic Acids Res.*, **41**, D43–D47.
2. Dimmer,E.C., Huntley,R.P., Alam-Faruque,Y., Sawford,T., O'Donovan,C., Martin,M.J., Bely,B., Browne,P., Mun Chan,W., Eberhardt,R. *et al.* (2012) The UniProt-GO Annotation database in 2011. *Nucleic Acids Res.*, **40**, D565–D570.
3. Yu,N.Y., Laird,M.R., Spencer,C. and Brinkman,F.S. (2011) PSORTdb–an expanded, auto-updated, user-friendly protein subcellular localization database for Bacteria and Archaea. *Nucleic Acids Res.*, **39**, D241–D244.
4. Briesemeister,S., Rahnenfuhrer,J. and Kohlbacher,O. (2010) Going from where to why–interpretable prediction of protein subcellular localization. *Bioinformatics*, **26**, 1232–1238.
5. Mei,S. (2012) Predicting plant protein subcellular multi-localization by Chou's PseAAC formulation based multi-label homolog knowledge transfer learning. *J. Theor. Biol.*, **310**, 80–87.

6. Habib,T., Zhang,C., Yang,J.Y., Yang,M.Q. and Deng,Y. (2008) Supervised learning method for the prediction of subcellular localization of proteins using amino acid and amino acid pair composition. *BMC Genomics*, **9(Suppl. 1)**, S16.
7. Hoglund,A., Donnes,P., Blum,T., Adolph,H.W. and Kohlbacher,O. (2006) MultiLoc: prediction of protein subcellular localization using N-terminal targeting sequences, sequence motifs and amino acid composition. *Bioinformatics*, **22**, 1158–1165.
8. Sarda,D., Chua,G.H., Li,K.B. and Krishnan,A. (2005) pSLIP: SVM based protein subcellular localization prediction using multiple physicochemical properties. *BMC Bioinformatics*, **6**, 152.
9. Xie,D., Li,A., Wang,M., Fan,Z. and Feng,H. (2005) LOCSVMPSI: a web server for subcellular localization of eukaryotic proteins using SVM and profile of PSI-BLAST. *Nucleic Acids Res.*, **33**, W105–W110.
10. Bhasin,M. and Raghava,G.P. (2004) ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. *Nucleic Acids Res.*, **32**, W414–W419.
11. Petsalaki,E.I., Bagos,P.G., Litou,Z.I. and Hamodrakas,S.J. (2006) PredSL: a tool for the N-terminal sequence-based prediction of protein subcellular localization. *Genomics Proteomics Bioinformatics*, **4**, 48–55.
12. Shi,J.Y., Zhang,S.W., Pan,Q., Cheng,Y.M. and Xie,J. (2007) Prediction of protein subcellular localization by support vector machines using multi-scale energy and pseudo amino acid composition. *Amino Acids*, **33**, 69–74.
13. Blum,T., Briesemeister,S. and Kohlbacher,O. (2009) MultiLoc2: integrating phylogeny and Gene Ontology terms improves subcellular protein localization prediction. *BMC Bioinformatics*, **10**, 274.
14. Anurag,M., Singh,G.P. and Dash,D. (2012) Location of disorder in coiled coil proteins is influenced by its biological role and subcellular localization: a GO-based study on human proteome. *Mol. Biosyst.*, **8**, 346–352.
15. Fyshe,A., Liu,Y., Szafron,D., Greiner,R. and Lu,P. (2008) Improving subcellular localization prediction using text classification and the gene ontology. *Bioinformatics*, **24**, 2512–2517.
16. Briesemeister,S., Blum,T., Brady,S., Lam,Y., Kohlbacher,O. and Shatkay,H. (2009) SherLoc2: a high-accuracy hybrid method for predicting subcellular localization of proteins. *J. Proteome Res.*, **8**, 5363–5366.
17. Chi,S.M. and Nam,D. (2012) WegoLoc: accurate prediction of protein subcellular localization using weighted Gene Ontology terms. *Bioinformatics*, **28**, 1028–1030.
18. King,B.R., Vural,S., Pandey,S., Barteau,A. and Guda,C. (2012) ngLOC: software and web server for predicting protein subcellular localization in prokaryotes and eukaryotes. *BMC Res. Notes*, **5**, 351.
19. Wan,S., Mak,M.W. and Kung,S.Y. (2012) mGOASVM: multi-label protein subcellular localization based on gene ontology and support vector machines. *BMC Bioinformatics*, **13**, 290.
20. Briesemeister,S., Rahnenfuhrer,J. and Kohlbacher,O. (2010) YLoc–an interpretable web server for predicting subcellular localization. *Nucleic Acids Res.*, **38**, W497–W502.
21. Wu,Z.C., Xiao,X. and Chou,K.C. (2011) iLoc-Plant: a multi-label classifier for predicting the subcellular localization of plant proteins with both single and multiple sites. *Mol. Biosyst.*, **7**, 3287–3297.
22. Chou,K.C. and Shen,H.B. (2010) Plant-mPLoc: a top-down strategy to augment the power for predicting plant protein subcellular localization. *PLoS One*, **5**, e11335.
23. Chou,K.C. and Shen,H.B. (2007) Large-scale plant protein subcellular location prediction. *J. Cell. Biochem.*, **100**, 665–678.
24. Lee,D.W., Kim,J.K., Lee,S., Choi,S., Kim,S. and Hwang,I. (2008) Arabidopsis nuclear-encoded plastid transit peptides contain multiple sequence subgroups with distinctive chloroplast-targeting sequence motifs. *Plant Cell*, **20**, 1603–1622.
25. Hu,Y., Li,T., Sun,J., Tang,S., Xiong,W., Li,D., Chen,G. and Cong,P. (2012) Predicting gram-positive bacterial protein subcellular localization based on localization motifs. *J. Theor. Biol.*, **308**, 135–140.
26. Tang,S.N., Sun,J.M., Xiong,W.W., Cong,P.S. and Li,T.H. (2012) Identification of the subcellular localization of mycobacterial proteins using localization motifs. *Biochimie*, **94**, 847–853.
27. Horton,P., Park,K.J., Obayashi,T., Fujita,N., Harada,H., Adams-Collier,C.J. and Nakai,K. (2007) WoLF PSORT: protein localization predictor. *Nucleic Acids Res.*, **35**, W585–W587.
28. Emanuelsson,O., Nielsen,H., Brunak,S. and von Heijne,G. (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.*, **300**, 1005–1016.
29. Emanuelsson,O., Brunak,S., von Heijne,G. and Nielsen,H. (2007) Locating proteins in the cell using TargetP, SignalP and related tools. *Nat. Protoc.*, **2**, 953–971.
30. Bannai,H., Tamada,Y., Maruyama,O., Nakai,K. and Miyano,S. (2002) Extensive feature detection of N-terminal protein sorting signals. *Bioinformatics*, **18**, 298–305.
31. Xiong,W., Li,T., Chen,K. and Tang,K. (2009) Local combinational variables: an approach used in DNA-binding helix-turn-helix motif prediction with sequence information. *Nucleic Acids Res.*, **37**, 5632–5640.
32. Sun,J., Tang,S., Xiong,W., Cong,P. and Li,T. (2012) DSP: a protein shape string and its profile prediction server. *Nucleic Acids Res.*, **40**, W298–W302.